# Root Cause Analysis of INC000000032544
# Open Science Grid: OSG Display Data Issue
***Report submitted 04/29/2010***
***(CD-doc-3928-v1)***

## Summary of Incident

OSG Display provides a high-level view of OSG activities -- graphs of jobs and data transfer statistics from the last several hours extracted from Gratia accounting system. Display home: http://display.grid.iu.edu. On Tuesday 04/20/2010 at about 10:00 AM CDT it was noticed that OSG display showed data that were last updated about 16 hours back. Brian Bockelman conducted initial investigation and noticed that some of the queries sent by OSG Display updater have not been completing within 5 minute window resulting in stale data shown in the OSG Display, and, by looking at Ganglia monitoring plot, he observed that db replication between "collector" and "reporter" databases (set up as Master/Slave pair) was 16 hours behind and the backlog had been building for the last 16 hours, resulting in absence of up to date data in "reporter" database. The latter is queried by OSG Display. Queries, similar to those executed by OSG Display, when run by hand, were not completing in reasonable (usual) time, indicating issue on the "reporter" database. Brian sent message to this effect via IM to Philippe Canal of Fermilab at about 10:00 AM CDT. Subsequently at about 10:30 AM CDT Chris Green followed the established procedure and failed over "reporter" database to "collector" database by switching service IP of the "collector" database host to point at "reporter" database host IP. At this point FermiGrid group considered incident resolved, but OSG Display have not been updating information until about 1:00 PM CDT b/c data collection on the OSG Display was disabled in the process of debugging the incident. It started to display up to date plots after that time.

The issue was e-mailed to OSG Facility mailing list and later escalated on the OSG side by opening GOC ticket #8456 and reported as "Gratia outage" at the OSG production meeting.

Tickets opened:
- GOC ticket #8456 (https://ticket.grid.iu.edu/goc/viewer?id=8456)
- Fermilab Service Desk ticket INC000000032544
- Fermilab Service Desk problem ID: PBI000000000135

## Background Concepts

### OSG Display

According to information available on OSG site ( https://twiki.grid.iu.edu/bin/view/MeasurementsAndMetrics/OSGDisplay) the "OSG Display is currently in alpha / draft form and is liable to change" (as of 04/06/2010). Based on information available on the web, the application seems to consist of two pieces – data display and updater. Updater executes queries by connecting directly to Gratia "reporter" database every 5 minutes. After queries complete the update of display is triggered. The following time series information is displayed:
- #jobs/hour during the last 24 hours
- TBytes/hour of data transferred during the last 24 hours
- #jobs/month during the last 12 months

- #transfers/month during the last 12 months.

The display has look and feel of application that provides up to date (with precision of 5 minutes) information about OSG jobs.

## Support model

The OSG Display service is in alpha state and no SLA agreement exist to support it.

## Gratia

Gratia is OSG and FermiGrid usage accounting system. It consists of multiple probes, collectors and reporters web services, and DB backend. The probes run on OSG sites and send metrics embedded in XML messages to collector processes. Collectors log data to "collector" database both in "raw" XML format as well as structured via XML->Java Object->Hibernate->Relation chain. The "collector" database is replicated to "reporter" database. Reporters provide r/o access to "reporter" database. There are 4 collector and corresponding reporter services:

- gratia-osg-prod
- gratia-osg-daily
- gratia-osg-transfer
- gratia-osg-itb

By design Gratia guarantees storage of all data sent in by the probes. As probes scheduling may differ from site to site the accounting information cannot be guaranteed to be up to date.

Besides providing summary data, the Gratia system records job level information (file transfers per job), and captures all incoming raw XML data from probes. Recording file transfers is high frequency, high load activity with gratia-osg-transfer generating up to 5M inserts per day in the "collector" DB. In order to cope with fast growing database size the fine granularity data is removed periodically by housekeeping processes that execute queries that remove data older than present number of days.

## Gratia deployment @ Fermilab

The Gratia system at FNAL was recently upgraded to increase the headroom and capability of the system. As part of the upgrade, the system was split onto two servers with a new replication model as described below.

Gratia system at Fermilab is deployed on two powerful nodes gratia12, and gratia13. Each of the 4 collector services runs in separate VM hosted by gratia12, the MySQL "collector" DB runs on 5$^{th}$ separate VM. Reporter services arranged in similar way on gratia13.

MySQL backend uses innoDB storage engine which provides transaction support and foreign keys (features missingg in vanilla MySQL distribution). "Collector" DB is write only database. A Master/Slave replication is set between "collector" db and "reporter" db.

VMs hosting databases have service IPs assigned to them. A failover scenario exists to use "collector" database instead of "reporter" database if the latter is rendered disabled by switching service IP so that reporter services will continue to function w/o reconfiguration. Failover is not automatic to prevent

run-away queries from killing “reporter” database and then killing “collector” database resulting in data loss due to inability of collector processes to log their data.

Fine granularity data, like job level information, as well as raw XML data are periodically removed from “collector” database keeping only data gathered during last 3 months (1 month for raw XML data).

At the moment of incident Gratia used MySQL v5.0.77 as patched by RedHat to 5.0.77-4.el5_4.2.

### Support model

Gratia is operated by FermiGrid group as production service on 8x5 basis. There is no SLA for Gratia between FermiGrid and OSG.

## Timeline

| | |
|---|---|
| 04/20/2010 | |
| 10:00:00 AM CDT | Brian Bockelman sends IM message to Phillipe Canal indicating that OSG Display queries take longer than usual to execute resulting in OSG Display not updating and that db replication has fallen behind. |
| 10:30:00 AM CDT | Chris Green executes manual failover by pointing service IP of “reporter” DB host to “collector” DB host |
| 01:35:00 PM CDT | Ruth Pordes opens GOC ticket #8456 (https://ticket.grid.iu.edu/goc/viewer?id=8456 ) |
| 01:45:00 PM CDT | Ticket INC000000032544 is created in Fermilab Service Desk. Urgency ”Medium”. Assigned to Gratia Operations |
| 02:57:00 PM CDT | Work log entry in INC000000032544 indicates that after “reporter” DB failed over to “collector” DB OSG display is getting updated data. Closed by GOC. |
| ~ 3:00 PM CDT | Brian reported on the production call that there had been an outage of the Gratia DB and as a result, “the OSG Display was unable to update for ~20 hours”.<br><br>Notes from the Production meeting, written by Dan F, were sent out to the OSG Facilities distribution list including the OSG Executive team. It was noted that there was a 20 hour Gratia outage, with no report from the Gratia team indicating a problem and there were plans to follow up. (http://twiki.grid.iu.edu/bin/view/Production/Apr20%2c2010ProductionMeeting) |
| 03:04:00 PM CDT | Ticket is reopened per Ruth Pordes' request. |
| 04/21/2010 | |
| 06:11:00 AM CDT | Adding Rob to ticket so he is aware of this discussion. Ruth, I think you already know this -- OSG_Display is still in semi-production state, with no SLA as far as I know. We still keep an eye on its status but not at a significant level like we do with BDII, MyOSG, etc.<br><br>Cheers,<br>-ag<br><br>For complete ticket details, click |

https://ticket.grid.iu.edu/goc/viewer?id=8456

03:07:00 PM CDT Ticket closed

## Analysis

A Root Cause Analysis meeting was held on 04/29/2010 with representatives of stakeholders (OSG) and service providers (FermiGrid group). Below is the list of people who attended the meeting with their roles indicated:

Brian Bockelman (Nebraska, OSG Display architect)
Phillipe Canal (Fermilab, Gratia architect and developer)
Keith Chadwick (Fermilab, FermiGrid operations project leader)
Dan Fraser (ANL, OSG Production Coordinator)
Chris Green (Fermilab, Gratia developer and operator, OSG User Group support)
Dmitry Litvintsev (Fermilab, problem coordinator, chair)
Rob Quick (Indiana University, OSG Operations Coordinator)
Dan Yocum (Fermilab, FermiGrid application support).

The chain of events during incident, environment and result of preliminary investigation were discussed in some detail. Main causes and possible causes based on this discussion were identified and are presented in a form of fishbone diagram in Figure 1. The diagram was made after analysis of information gathered during RCA meeting.

The causes that may have contributed to the occurrence of incident "OSG Display shows stale data" belong roughly to three categories – environmental , technical (technology in the graph) and operational.

**Technology**

- OSG Display shows stale data:
  - OSG Display updater is not updating data shown by display.
  - Queries executed by OSG updater on Gratia "reporter" db do not complete in 5 minute window.
  - Queries are slow b/c due to bad query plan ignoring index and favoring full table scan.
  - Other queries running on "reporter" db, like housekeeping queries, suffer from the same problem.
  - Queries running full table scans contribute to "reporter" db slowdown.
  - Bad query plan is due to miscalculation of index cardinality.
  - Gratia uses MySQL v5.0.0.77 as patched by RedHat to 5.0.77-4.el5_4.2. It turns out that bad query plans were triggered by a bug in innoDB storage layer. Bug is acknowledged by MySQL, c.f.:
    - http://bugs.MySQL.com/bug.php?id=36513
    - http://bugs.MySQL.com/bug.php?id=43660

The bug in innoDB layer caused full table scans on large tables thrashing disk and therefore affecting all other processes on "reporter" db, including data replication from "collector" db. The data in

"reporter" db was 16 hours behind. If OSG Display queries had been returning within 5 minute window the data shown by OSG Display would still have been 16 hours behind, but this would not have been considered as alarm because Gratia is not expected to provide up to date data. The replication eventually caught up without intervention. Although, from OSG operations standpoint, after the upgrade mentioned above (with an order of magnitude improvement as discussed at the meeting), it had been anticipated by the OSG team that data would be reasonably current.

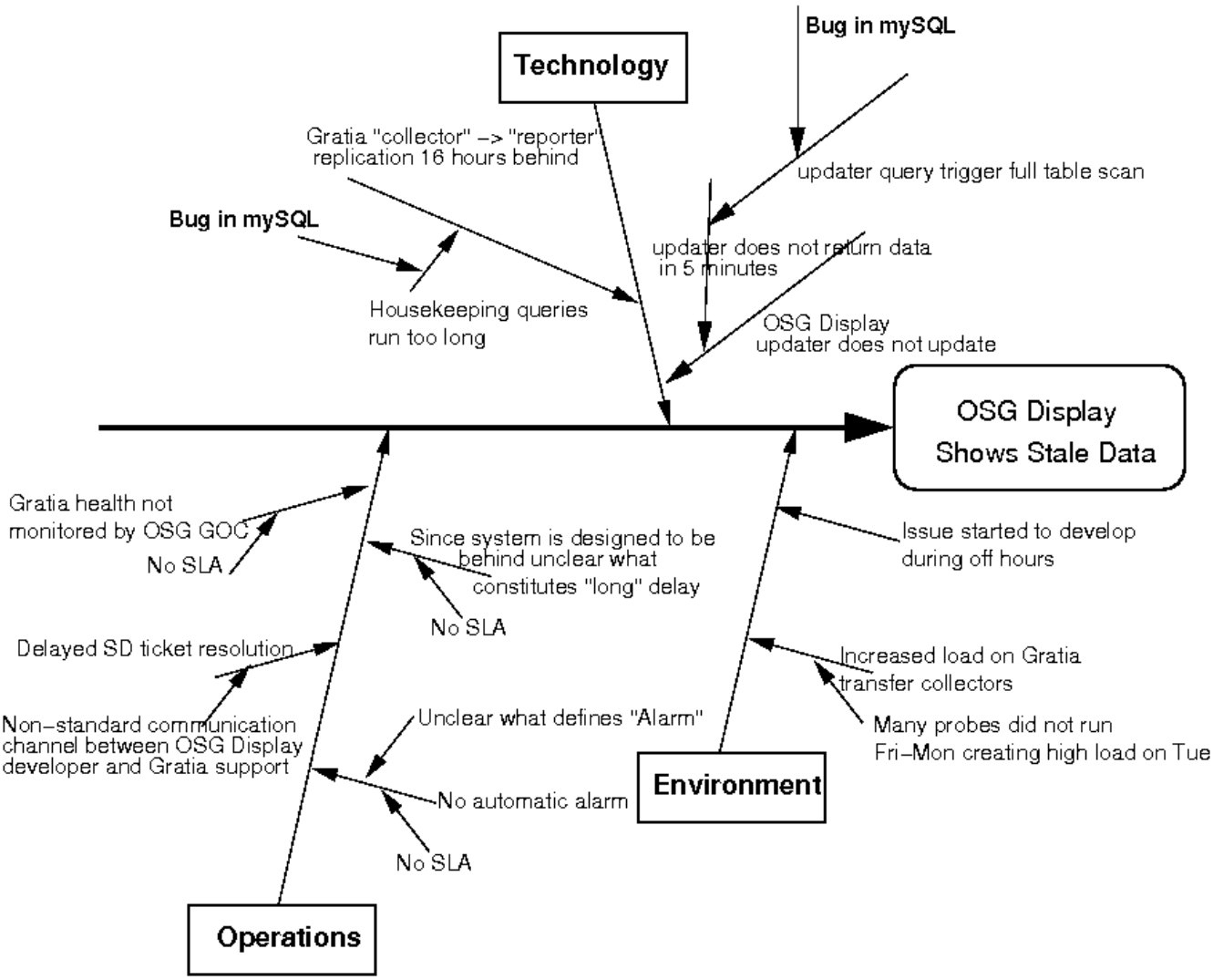


*Figure 1: Categorized caused that may have contributed to the incindent. The root cause of the incident is identified as "Bug in MySQL"*

Corrective action of failing over to "collector" db executed at 10:30 AM CDT went unnoticed by OSG Display because OSG Display updater was disabled till about 1:00 PM CDT as part of the investigation on OSG Display end.

The "collector" database was apparently not affected by the bug, and continued to log data provided by the probes and then serve OSG Display queries.

**Bug in innoDB storage layer of MySQL is direct cause of the incident.**

**Environment:**

- Prior to incident probes running on the FNAL T1 site were off on Fri-Mon. The accumulated data came in in large wave on Tuesday generating massive upload of job level data.
- "Collector" database was running nightly backup.
- The issue with OSG Display started to develop during off-business hours. This could have contributed to delayed reaction to the incident because it wasn't reported until start of the business day. Although, due to lack of SLA, it is unclear how the time it took to resolve the incident can be characterized (whether or not it was appropriate).

**Operational/procedural:**

- From OSG Display developer's point of view Gratia "reporter" database has developed an issue that resulted in slow down of queries sent by OSG Display to "reporter" database.
- Using non-standard communication channel the problem was reported to Gratia support.
- Problem was addressed withing 30 minutes after initial report by failing over to "collector" database at 10:30 AM CDT.
- No Service Desk ticket was created.
- It was not clear until later in the day (about 1:00 PM CDT) whether or not the corrective action solved OSG Display issue because the updater was disabled.
- Problem ticket was opened in GOC ticket system after the issue has been resolved. Therefore it was quickly closed resulting in corresponding ticket in Fermilab Remedy System (generated automatically when GOC ticket was created) being in "Cancelled" state preventing Fermigrid support from updating the ticket.
- Gratia system is not monitored by OSG Operations creating perception of lack of transparency from OSG clients point of view. The system is extensively monitored by FermiGrid, but this monitoring is not connected to OSG Operations monitoring pages.
- Absence of automated alert system that would create alarm to alert FermiGrid operations group or/and create ticket in Fermi Service Desk system shadowed in GOC ticket system may result in delayed reaction to Gratia incidents by FermiGrid operations team.
- It was unclear what constitutes alarm as system was not designed to provide up to date data.
- Existing monitoring did not show the type of slowdown registered by OSG Display because the "ping query" executed by monitoring system was not affected by MySQL bug.
- Lack of Gratia SLA creates situation where it is unclear what level of service is expected of Gratia and vice versa it is unclear what level of support needs to be provided by FermiGrid.

**Lack of Gratia SLA is considered an underlying problem that manifests itself in above listed operational/procedural issues.**

## Remedies implemented so far

- Configuration change was implemented on "reporter" database to always favor index scan on housekeeping queries. Same change was approved for "collector" database but the change requires some coordination as collectors need to be suspended during database restart.
- Internal alarm has been implemented to alert FermiGrid operations team that replication has fallen behind by a defined number of seconds. The following matrix of time threshold/alarm/action has been implemented:

| Time Threshold (seconds) | Alarm level | Action |
|---|---|---|
| 3600 | INFO | e-mail |
| 21600 | WARN | e-mail |
| 86400 | ERROR | e-mail |

- OSG is working with the Gratia team to see if alternative queries for the OSG Display may avoid hitting the MySQL bug and be better suited to the Gratia environment.

**Solutions:**

- Upgrade to MySQL v5.0.81 and db configuration change solves issue with the query explain plan. In the meantime the queries executed by OSG Display as well as some Gratia internal queries need to be revisited such that index scan is always forced.
- OSG Display running on http://display.grid.iu.edu would benefit from an annotation similar to this : "The data shown are obtained by querying Gratia system every 5 minutes. Check Gratia requirements document on how the data are collected" (underlined text implies hyperlink). A sort of fine print statement.
- SLA on Gratia has to be made between FermiGrid and OSG.
  - As service requirements will be defined and agreed, a metrics that shows Gratia system status will be made available to OSG GOC.
  - OSG is requesting that agreed upon metrics be transparent and machine readable for inclusion into OSG service monitoring available on the web.
  - Service requirements will determine alarm levels that will be implemented to automatically alert FermiGrid support group and/or open tickets in Fermilab Service Desk system and possibly in GOC ticket system so that status of Gratia is transparent to both clients and service providers.
- FermiGrid operations should provide regular reports and attend weekly OSG operations meetings to discuss the operation of Gratia with OSG staff.
- As OSG Display becomes Production an SLA covering its interaction with Gratia should be drawn.

**Appendix:**

Gratia Service Monitor:
http://fermigrid.fnal.gov/monitor/fermigrid0-gratia-service-summary.html

Ganglia Monitoring:

http://fg2x2.fnal.gov/ganglia/?r=week&c=FermiGrid&h=gr13x5.fnal.gov

http://fg3x2.fnal.gov/ganglia/?r=week&c=FermiGrid&h=gr13x5.fnal.gov

FermiGrid Service Monitor:

http://fermigrid.fnal.gov/fermigrid-metrics.html

An example of existing monitoring plots showing how replication is monitored.

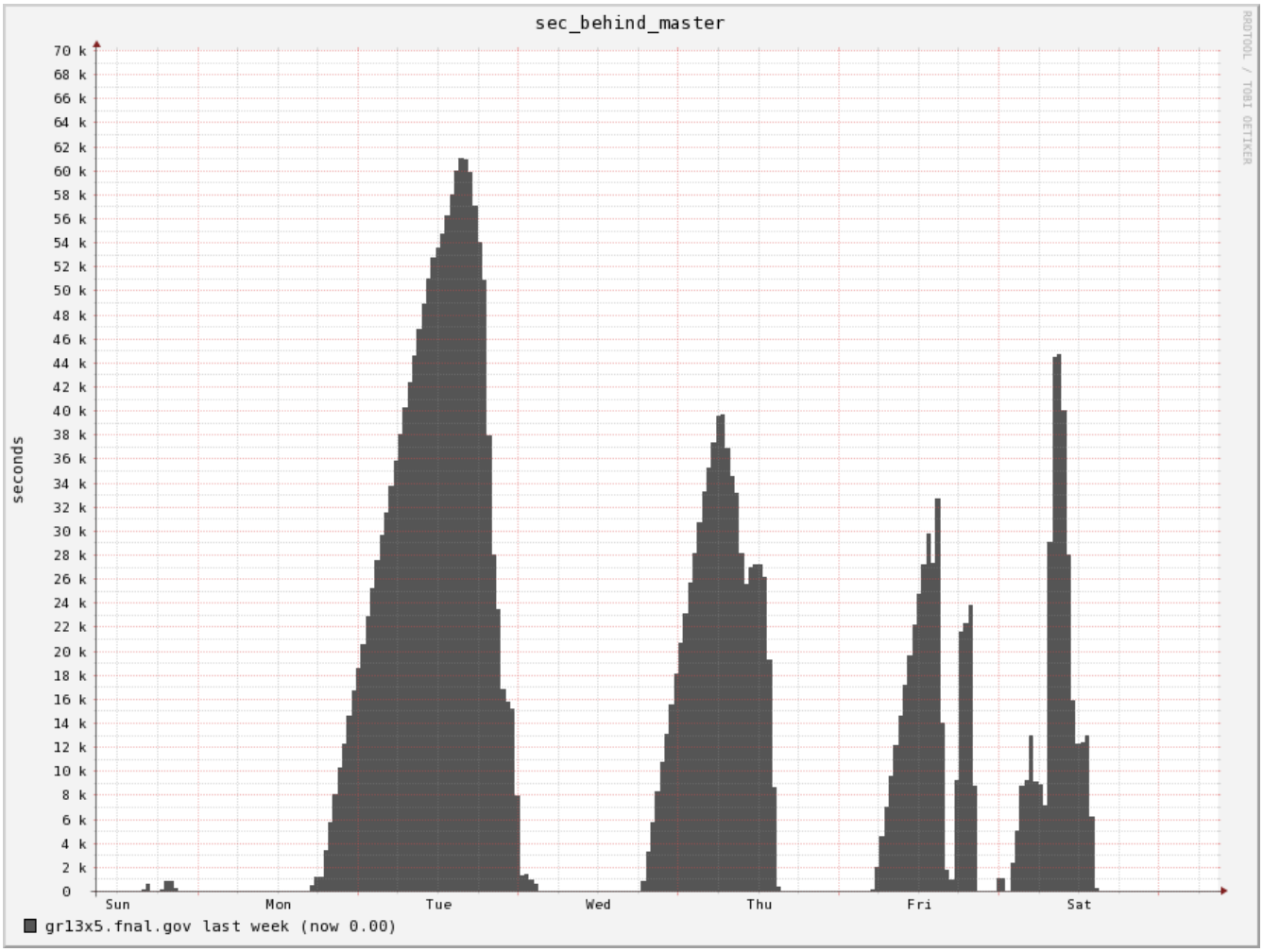


*Figure 2: Number of seconds the replication is behind vs time. Incident happened on Tuesday.*